AML5103 | Applied Probability and Statistics | Problem Set-2

1. You have tracked the performance of the local meteorologist and complied the following data:

   P(forecast rain, and actual rain) = 0.4, P(forecast rain, and no rain) = 0.2,
   P(forecast no rain, and actual rain) = 0.15, P(forecast no rain, and no rain) = 0.25.

   (a) How often does she forecast rain?

   (b) How often does she make a mistake?

   (c) Given that she just forecast rain, what is the chance that it will actually rain?

   (d) Given that it rains today, what is the probability that she had forecast rain in last night's broadcast?

2. Fifty-two percent of the students at a certain college are females. Five percent of the students in this college are majoring in computer science. Two percent of the students are women majoring in computer science. If a student is selected at random, find the conditional probability that
   (a) the student is female given that the student is majoring in computer science;
   (b) this student is majoring in computer science given that the student is female.

3. Suppose that an insurance company classifies people who buy medical insurance from them into one of three classes: good, average, and bad risks. As a data scientist for the company, you have access to the following customer data for the calendar year 2023-24:

Class	**% of customers**	**% who had a major health problem**
Bad risk	20%	25%
Average risk	40%	20%
Good risk	40%	10%

   What is the probability that a new customer will not meet with a major health problem during 2024-25?

4. Suppose there are four failure modes for a single engine plane: *structural, engine, control system, and human error*. We assume that they will only occur one at a time if at all.

The probabilities for these four failure modes are respectively 0.002, 0.002, 0.01, and 0.001. Given that it is a structural failure, there is a 25% chance the plane will crash. The crash probabilities given the other three failure modes are 30%, 90% and 10% respectively for engine, control system and human error failure modes. If a plane has crashed, what is the probability that it was due to a control system failure?

5. A robot, which only has a camera as a sensor, can either be in one of two locations: L_1 or L_2. The robot doesn't know exactly where it is but based on all past observations, the robot thinks that there is an 80% chance that it is in L_1 and a 20% chance that it is in L_2. Location L_2 is the only one that has a window. The robot's vision algorithm detects a window but its image recognition algorithm is not perfect; the probability of observing a window given there is no window at its location is 0.2 and the probability of observing a window given there is a window is 0.9. After incorporating the observation of a window, what is the robot's updated probability that it is in (1) L_1 (2) L_2?

6. Suppose a person needs a blood transfusion. We show the compatibility chart (where an X means compatibility) and population blood type percentage below (Source: Palo Alto Weekly, November 9, 1994):

Population Percentage	
O-	6.6%
O+	37.4%
A-	6.3%
A+	35.7%
B-	1.5%
B+	8.5%
AB-	0.6%
AB+	3.4%

Recipient ↓ / Donor →	O-	O+	A-	A+	B-	B+	AB-	AB+
O-	X							
O+	X	X						
A-	X		X					
A+	X	X	X	X				
B-	X				X			
B+	X	X			X	X		
AB-	X		X		X		X	
AB+	X	X	X	X	X	X	X	X

Compatibility Chart

Given the above information, examine the following questions. Also, pose other meaningful questions regarding the supply and demand of blood (types) given the above compatibility table and blood type distribution.

(a) What is the probability that a random person will be able to donate to another random person given no information about blood types of either the giver or the receiver?

(b) Given the above data, what can you say about blood transfusion policy in a hospital regarding blood drives target and blood transfusion priority?

(c) In a battle field hospital a soldier is brought in for immediate blood transfusion. Only blood type A+ is available in the supply for immediate use. We do not know the wounded soldier's blood type. There are two other soldiers present who are willing to donate their blood. We have time to do one blood typing before time becomes critical. What should we do?

7. In a future society, a machine is used to predict a crime before it occurs. If you were responsible for tuning this machine, what evaluation metric would you want to maximize to ensure no innocent people (people not about to commit a crime) are imprisoned?

8. Consider a classification model that separates email into two categories: "spam" or "not spam." Answer the following questions regarding precision and recall (a.k.a. sensitivity or true positive rate) by playing around with the threshold slider on the demo website here:
   (a) Which is a more relevant performance metric in this case: recall or precision? Explain briefly why.
   (b) Increasing the classification threshold generally $\underbrace{\text{increases/decreases}}_{\text{choose one}}$ FP.
   (c) When the classification threshold increases, precision
   $$\underbrace{\text{probably increases/probably decreases/definitely increases/definitely decreases}}_{\text{choose one}}.$$
   (d) Keeping in mind that $TP + FP + TN + FN = n$, which is the number of samples, when the classification threshold is increased, what happens to the quantity TP?
   (e) When the classification threshold is increased, the quantities TN and FN both
   $$\underbrace{\text{uniformly/non-uniformly}}_{\text{choose one}} \ \underbrace{\text{increase/decrease}}_{\text{choose one}}.$$

(f) Decreasing the classification threshold generally $\underbrace{\text{increases/decreases}}_{\text{choose one}}$ FN.

(g) When the classification threshold is decreased, recall

$$\underbrace{\text{probably increases/probably decreases/definitely increases/definitely decreases}}_{\text{choose one}}.$$

(h) When the classification threshold is decreased, the quantities TP and FP both

$$\underbrace{\text{uniformly/non-uniformly}}_{\text{choose one}} \ \underbrace{\text{increase/decrease}}_{\text{choose one}}.$$

9. In which of the following scenarios would a high accuracy value suggest that the ML model is doing a good job? Explain your answer briefly.

   (a) An expensive and critical hydro-electric turbine operates 23 hours a day. An ML model evaluates vibration patterns and predicts when the turbine is operating without anomaly with an accuracy 99.99%.

   (b) You are building an ML tool for a retail company which will predict, based on past purchase history and other demographic information, the high end cellphone that the next buyer will potentially buy from an available 10 high end models. Your ML model has an accuracy of 15%.

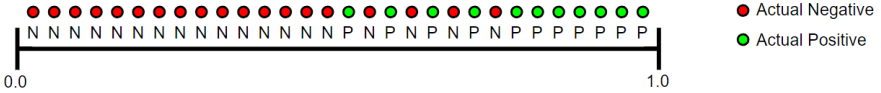
(c) A deadly, but curable, medical condition afflicts .01% of the population. Your ML model uses symptoms as features and predicts this affliction with an accuracy of 99.99%.

10. Consider two models: A and B, that each evaluate the same dataset. Which one of the following statements is true?
    (a) If model A has better precision and better recall than model B, then model A is probably better.
    (b) If model A has better recall than model B, then model A is better.
    (c) If Model A has better precision than model B, then model A is better.

11. An ROC curve is a plot of ____ vs. ____ for different _____.

12. Lowering the classification threshold classifies more items as $\underbrace{\text{positive/negative}}_{\text{choose one}}$, thus increasing both _____ and _____.

13. AUC (Area under the ROC Curve) provides an aggregate measure of performance across all possible classification thresholds. One way of interpreting AUC is as the probability that the model ranks a random positive example more highly than a random negative example. For example, given the following examples, which are arranged from left to right in ascending order of prediction probabilities:



AUC represents the probability that a random positive (green) example is positioned to the right of a random negative (red) example. AUC ranges in value from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0.0; one whose predictions are 100% correct has an AUC of 1.0.

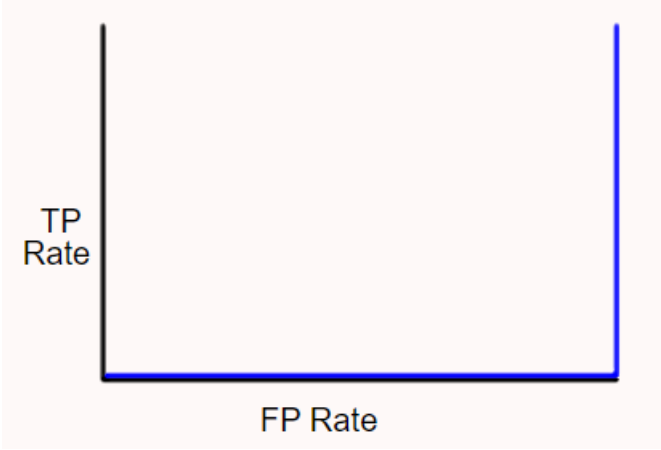
Suppose we multiplied all of the predictions from a given model by 0.5 (for example, if the model predicts 0.4, we multiply by 0.5 to get a prediction of 0.2), how would it change the model's performance as measured by AUC?

(a) It would make AUC terrible, since the prediction values are now way off.

(b) It would make AUC better, because the prediction values are all farther apart.

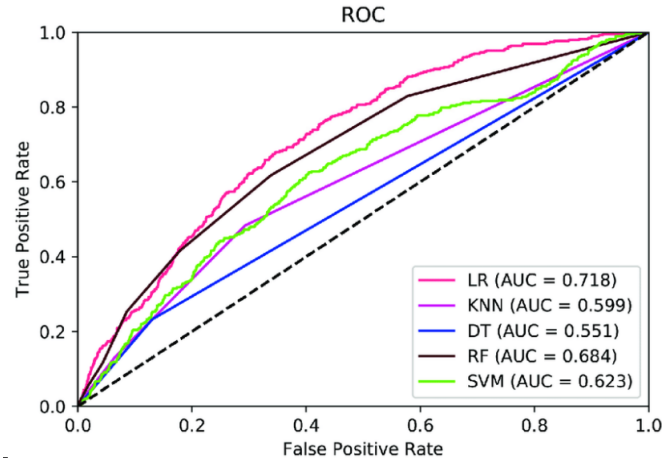
(c) No change. AUC only cares about relative prediction probabilities.

14. Your friend shows you his model's ROC curve as follows:



Is your friend's model any good, why? How can you help your friends model go from zero to hero?

15. The figure shows ROC curves for different models.

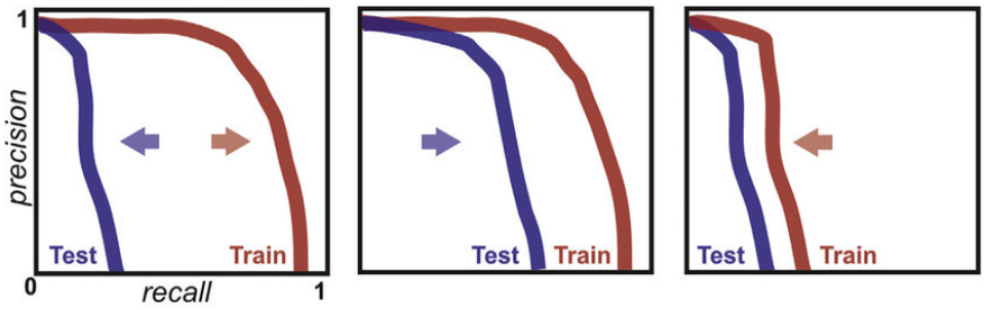


Classify the following statements as true or false:

- Dashed black line represents random classification.
- ROC curve for any model can't fall below the dashed black line.
- The model represented by solid blue line is better than that represent by solid lime.

16. Which one among TP, TN, FP, FN does not play a role in forming the precision-recall curve? What does the conclusion mean intuitively?

17. The figure shows precision-recall curves for different models on the train and test sets.



Identify which model overfits, which one underfits, and which one is a good fit.

18. Explain which one among area under ROC and area under precision-recall curve would you use for the following scenarios:
    (a) Identifying whether a customer will buy a product on discount or not when a customer is equally likely to do so.
    (b) Identifying a spam email when generally spam emails constitute 1% of the total emails.